МИНИСТЕРСТВО ОБРАЗОВАНИЯ РЕСПУБЛИКИ БЕЛАРУСЬ

БЕЛОРУССКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

ФАКУЛЬТЕТ РАДИОФИЗИКИ И КОМПЬЮТЕРНЫХ ТЕХНОЛОГИЙ

Кафедра системного анализа и компьютерного моделирования

СИКОЛЕНКО Михаил Александрович

АВТОМАТИЗИРОВАННАЯ АННОТАЦИЯ ТРАНСКРИПТОВ В ГЕНОМАХ ПРОКАРИОТ С ИСПОЛЬЗОВАНИЕМ СВЁРТОЧНЫХ НЕЙРОННЫХ СЕТЕЙ

Аннотация (реферат) к дипломной работе

Научный руководитель:
кандидат физико-математических наук,
доцент В.В. Скакун

Допущена к защите
«___» _____ 2025 г.
Зав. кафедрой системного анализа и компьютерного моделирования

кандидат физико-математических наук, доцент _____ В.В. Скакун

Минск, 2025

РЕФЕРАТ

В дипломной работе 56 страницы, 32 рисунков, 3 таблиц, 35 источника, 1 приложение.

Ключевые слова: автоматизированная аннотация транскриптов, прокариотические геномы, свёрточные нейронные сети, глубокое обучение, машинное обучение, RNA-Seq, биоинформатика, анализ последовательностей, SHAP-интерпретация, конвейер обработки данных

Аннотация транскриптов прокариотических геномов крайне важна для понимания механизмов регуляции экспрессии генов. Существующие конвейеры (PGAP, SigmoID) ограничены анализом только нкРНК, тРНК и рРНК, в то время как предсказание координат транскриптов *de novo* остаётся нерешённой задачей.

Цель работы – разработать и реализовать метод глубокого и обучения для автоматизированного предсказания координат транскриптов в прокариотических геномах исключительно на основе последовательности ДНК, без использования RNA-Seq данных.

Объект исследования – биофизические и последовательностные признаки транскриптов в геномах прокариот.

Предмет исследования – алгоритмы и методы машинного и глубокого обучения для автоматизированного предсказания координат транскриптов по последовательности ДНК.

Методика включала построение программного конвейера: индексирование генома (Entrez, Bowtie 2), очистка и QC-анализ «сырых» прочтений (Fastp, FastQC), картирование (Bowtie 2/SAMtools) и реконструкцию StringTie с разбиением генома на перекрывающиеся окна 5 376 пн (шаг 960) и маркировкой по покрытию ($> 66\,\%/\leq 66\,\%$). Для каждого окна рассчитывались частоты *k*-меров, энтропийные и спектральные характеристики, а также параметры открытых рамок считывания; отбор информативных признаков проводился через корреляционный анализ, дисперсионный анализ и на основании вклада призаков в модель «случайный лес».

Свёрточная нейронная сеть показала метрику качества ROC-AUC равную 0,969, а классические алгоритмы машинного обуечняи настраивались с помощью библиотеки PyCaret. Ансамбль из свёрточной нейронной сети и метода K-ближайших соседей обеспечил метрику качества F1-меру равную 0,964.

В результате был разработан программный пакет, включающий алгоритмы для полного конвейера аннотации на языке оболочки BASH, JUPYTER-ноутбуки с реализацией моделей и PYTHON-модули, формирующие GTF-файл транскриптов без использования RNA-Seq данных.

РЭФЕРАТ

У дыпломнай рабоце 56 старонкі, 32 ілюстрацыі, 3 табліцы, 35 крыніцы, 1 дадатак.

Ключавыя словы: аўтаматызаваная анатацыя транскрыптаў, пракарыётычныя геномы, згорткавыя нейронныя сеткі, глыбокае навучанне, машыннае навучанне, RNA-Seq, біяінфарматыка, аналіз паслядоўнасцей, SHAP-інтэрпрэтацыя, канвеер апрацоўкі данных.

Анатацыя транскрыптаў пракарыётычных геномаў надзвычай важная для разумення механізмаў рэгуляцыі экспрэсіі генаў. Існуючыя канвееры (PGAP, SigmoID) абмяжоўваюцца аналізам толькі нкРНК, тРНК і рРНК, у той час як *de novo* прагназаванне каардынат транскрыптаў застаецца невырашэнная задачай.

Мэта работы – распрацаваць і рэалізаваць метад глыбокага і машыннага навучання для аўтаматызаванага прагназавання каардынат транскрыптаў у пракарыётычных геномах выключна на аснове паслядоўнасці ДНК, без выкарыстання даных RNA-Seq.

Аб’ект даследавання – біфізічныя і паслядоўнасныя прыкметы транскрыптаў у геномах пракарыёт.

Прадмет даследавання – алгарытмы і метады машыннага і глыбокага навучання для аўтаматызаванага прагназавання каардынат транскрыптаў па паслядоўнасці ДНК.

Методыка ўключала пастраенне праграмнага канвеера: індэксаванне генома (Entrez, Bowtie 2), ачыстка і QC-аналіз «сырых» прачтенняў (Fastp, FastQC), картіраванне (Bowtie 2/SAMtools) і рэканструкцыю StringTie з разрэзам генома на перакрыжаваныя вокны па 5 376 пн (крок 960) і маркіроўкай па пакрыцці ($> 66\,\%/\leq 66\,\%$). Для кожнага вакна вылічваліся частоты *k*-мераў, энтрапійныя і спектральныя характарыстыкі, а таксама параметры адкрытых рамак счытвання; адбор інфарматыўных прыкмет праводзіўся праз карэляцыйны аналіз, дысперсійны аналіз і на аснове укладу прыкмет у мадэль «выпадковы лес».

Згорткавая нейронная сетка паказала метрыку якасці ROC-AUC, роўную 0,969, а класічныя алгарытмы машыннага навучання наладжваліся з дапамогай бібліятэкі PyCaret. Ансамбль са згорткавай нейроннай сеткі і метаду K-бліжэйшых суседзяў забяспечыў метрыку якасці F1-меру, роўную 0,964.

У выніку был распрацаваны праграмны пакет, які ўключае алгарытмы для поўнага канвеера анатацыі на мове абалонкі BASH, JUPYTER-ноутбукі з рэалізацыяй мадэляў і PYTHON-модулі, што фарміруюць GTF-файл транскрыптаў без выкарыстання дадзеных RNA-Seq.

STRUCTURAL ABSTRACT

The thesis comprises 56 pages, 32 figures, 3 tables, 35 references, and 1 appendix.

Keywords: automated transcript annotation, prokaryotic genomes, convolutional neural networks, deep learning, machine learning, RNA-Seq, bioinformatics, sequence analysis, SHAP interpretation, data-processing pipeline

Transcript annotation in prokaryotic genomes is critically important for understanding the mechanisms that regulate gene expression. Existing pipelines (PGAP, SigmoID) are limited to analysis of ncRNAs, tRNAs, and rRNAs, while de novo prediction of transcript coordinates remains an unsolved challenge.

The aim of this work is to develop and implement a method combining deep and classical machine learning for automated prediction of transcript coordinates in prokaryotic genomes based solely on DNA sequence, without using RNA-Seq data.

Object of study: biophysical and sequence-based features of transcripts in prokaryotic genomes.

Subject of study: machine-learning and deep-learning algorithms and methods for automated prediction of transcript coordinates from DNA sequence.

The methodology involved constructing a software pipeline: genome indexing (Entrez, Bowtie 2), cleaning and quality control of “raw” reads (Fastp, FastQC), mapping (Bowtie 2/SAMtools), and transcript reconstruction with StringTie after partitioning the genome into overlapping windows of 5 376 nt (step size 960) and labeling them by coverage ($> 66\,\%/\leq 66\,\%$). For each window, *k*-mer frequencies, entropy and spectral features, and open-reading-frame parameters were calculated; informative feature selection was performed via correlation analysis, analysis of variance, and feature-importance ranking in a random forest model.

A convolutional neural network achieved a ROC-AUC of 0.969, and classical machine-learning algorithms were tuned using the PyCaret library. An ensemble combining the CNN and the K-nearest neighbors method yielded an F1 score of 0.964.

As a result, a software package was developed, comprising a complete annotation pipeline implemented in BASH, JUPYTER notebooks with model implementations, and Python modules that generate a GTF-file of transcripts without relying on RNA-Seq data.